Modelo de regresión para la predicción de precios de automóviles

Data science essential training

DiplomadosOnline.com

Ejecute las siguientes lineas para continuar.

```
df = read.csv("data.csv", sep=",", header = TRUE)

library(janitor)
```

```
## Warning: package 'janitor' was built under R version 3.6.3
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
df = clean_names(df)

df$model <- gsub(" ", "_", df$model)
df$engine_fuel_type <- gsub(" ", "_", df$engine_fuel_type)
df$driven_wheels <- gsub(" ", "_", df$driven_wheels)

set.seed(1234)
trvaltest <- function(dat,prop = c(0.6,0.2,0.2)){
  nrw = nrow(dat)
  trnr = as.integer(nrw *prop[1])
  vlnr = as.integer(nrw*prop[2])
  set.seed(123)
  trni = sample(1:nrow(dat),trnr)
  trndata = dat[trni,]
  rmng = dat[-trni,]
  vlni = sample(1:nrow(rmng),vlnr)
  valdata = rmng[vlni,]
  tstdata = rmng[-vlni,]
  mylist = list("trn" = trndata,"val"= valdata,"tst" = tstdata)
  return(mylist)
}
outdata = trvaltest(df,prop = c(0.6,0.2,0.2))
df_train = outdata$trn; df_val = outdata$val; df_test = outdata$tst

library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
y_train_orig =  select(df_train, msrp)
y_val_orig = select(df_val, msrp)
y_test_orig = select(df_test, msrp)

y_train = log1p(y_train_orig)
y_val = log1p(y_val_orig)
y_test = log1p(y_test_orig)

df_train <- df_train[ ,!colnames(df_train)=="msrp"]
df_val <- df_val[ ,!colnames(df_val)=="msrp"]
df_test <- df_test[ ,!colnames(df_test)=="msrp"]
```

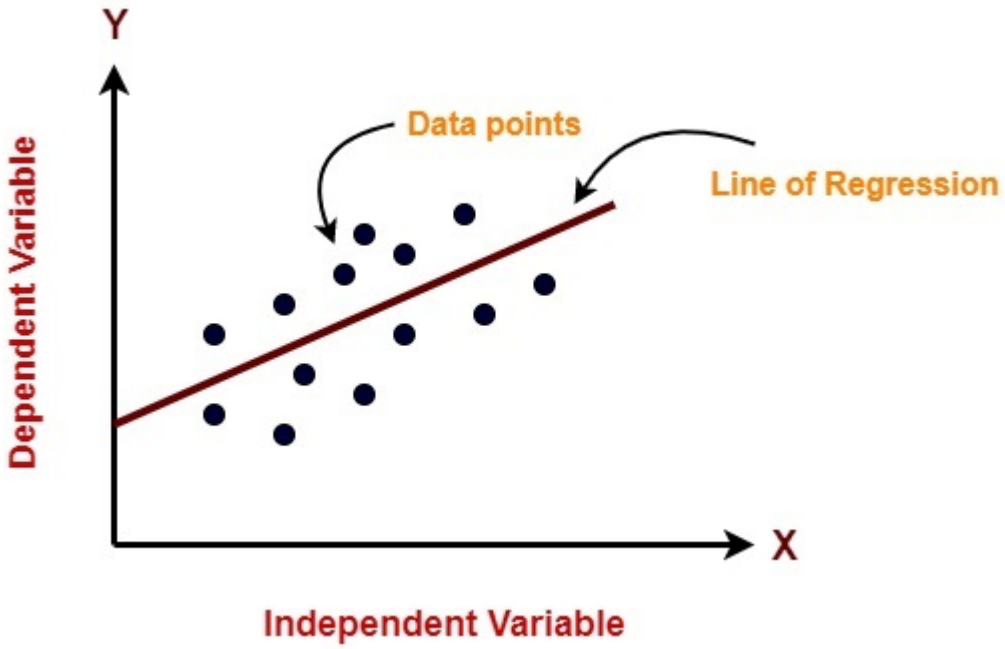
2) Modelado de los datos

Tras realizar el análisis inicial de los datos, estamos preparados para entrenar un modelo. El problema que estamos resolviendo es un problema de regresión: el objetivo es predecir un número, el precio de un automovil.

Para este proyecto utilizaremos el modelo de regresión más sencillo: la regresión lineal.

2.1 Regresión lineal

En primer lugar, repasemos cómo funciona la regresión lineal



Regresión lineal

Supongamos que tenemos una observación simple x_i y el valor y i que queremos predecir. El índice i significa aquí que se trata de la observación número i, una de las m observaciones que tenemos en nuestro conjunto de datos de entrenamiento.

Entonces, para esta única observación, el modelo a estimar tiene la forma:

[y_{i} f(x_{i})]

Si tenemos n características, nuestro vector x_i sería $n-$dimensional, por lo que tiene n componentes:

[x_{i} = (x_{i1},x_{i2},…, x_{in})]

Como tiene n componentes, podemos escribir la función f como una función con n parámetros, que es lo mismo que la fórmula anterior:

[y_{i} f(x_{i}) = f(x_{i1},x_{i2},…, x_{in})]

En nuestro caso, tenemos $7,150$ autos en el conjunto de datos de entrenamiento. Esto significa que $m = 7,150$ y i puede ser cualquier número entre 0 y 7,149.

Si f es el modelo de regresión lineal, tiene la siguiente forma:

[f(x_{i}) = f(x_{i1},x_{i2},…, x_{in})=*{0}*+{1}x_{i1}++*{n}*x{in}]

donde, $\beta_0, \ldots, \beta_n$ son los parámetros del modelo:

- β_0 es el término de sesgo
- $\beta_1, \ldots, \beta_n$ son las ponderaciones de cada característica $x_{i1}, x_{i2}, \ldots, x_{in}$.

Estos parámetros definen exactamente cómo debe combinar el modelo las características para que las predicciones al final sean lo mejor posible.

Para que la fórmula sea más corta, vamos a utilizar la notación de suma:

[f(x_{i}) = f(x_{i1},x_{i2},…, x_{in})=*{0}*+ {j=1}^{n}*{j}*x{ij}]

Entrenamiento:

¿Cuáles son x_i e y_i para este problema?

Estas ponderaciones son las que aprende el modelo cuando lo entrenamos.

Este modelo lo podemos escribir es forma matricial como:

$$f(X) = \beta_0 + X\beta,$$

donde, β es el vector columna de los coeficientes $\beta_1, \beta_2, \ldots, \beta_n$, mientras X es lo que se conoce como la matiz de diseño que contiene todas las observaciones de las características consideradas.

2.2 Entrenamiento del modelo de regresión

Para poder hacer prediciones, necesitamos saber las ponderaciones β. ¿Cómo las obtenemos?

Aprendemos los pesos a partir de los datos: utilizamos la variable objetivo y para encontrar la beta que combina las características de X de la mejor manera posible.

En el caso de la regresión lineal, “la mejor manera posible”significa que minimiza el error entre las predicciones $f(X)$ y el objetivo real y.

Tenemos varias formas de hacerlo. Utilizaremos la ecuación normal, que es el método más sencillo. El vector de pesos β se puede calcular con la siguiente fórmula:

$$\beta = (X^T X)^{-1} X^T y$$

Esto es fácil de traducir a NumPy:

- X^T es la transpuesta de X. En NumPy, es `t(X)`

- $X^T X$ es una multiplicación de matrices, que podemos hacer con el método del punto de NumPy: `t(X) %*% X`
- X^{-1} es la inversa de X. Podemos utilizar la función `solve()` para calcular la inversa.

Para implementar la ecuación normal, tenemos que hacer lo siguiente:

- Crear una función que tome una matriz X con características y un vector y con el objetivo.
- Añadir una columna ficticia (la característica que siempre se pone en 1) a la matriz X.
- Entrena el modelo: calcula los pesos β mediante la ecuación normal.
- Dividir este β en el sesgo β_0 y el resto de los pesos, y devolverlos.
- El último paso (dividir β en el término de sesgo y el resto) es opcional y se hace principalmente por conveniencia.

La implementación sería:

```
train_linear_regression = function(X,y){
    ones = as.vector(rep(1, dim(X)[1]))
    X = cbind(ones, X)

    XTX = t(X) %*% X
    XTX_inv = solve(XTX)
    w = XTX_inv %*% t(X) %*% y

    w
}
```

Entrenamiento:

Explique cada linea de la función e identifíquela con los pasos señalados.